Binary Classification Model With DL

Table of Contents

Objectives 2
Brief description of the data set and a summary of its attributes 2
Data Exploration 3
Data Cleaning and Feature engineering 4
Data Preparation for Machine Learning 5
  Target Variable 5
  Feature Selection 6
  Data Splitting 7
  Feature transformation 8
Model Training and Evaluation 8
Key Findings and Insights 11
Model Selection 18
Next Steps in model improvement 18
Summary 19

Objectives

The main objective of this report is to use the yellow taxicab dataset to build a binary classification model that can predict whether a trip will have a toll or not. Since toll cost adds to the total amount billed, this model will help the business to estimate the accurate total amount that a customer will be charged on any trip.

We will use different classification metrics to select the best model.

Brief description of the data set and a summary of its attributes

TLC Trip Record Data has 12 years (2009 –2020) worth of Data available but I have decided to analyze a subset of the 2015.

In 2015, passengers took nearly 300 million yellow cab rides in New York City. Working with the complete dataset for all these rides would require considerable time and computational resources. The 12 data files used represent two percent of the total trips sampled at random from each month.

I chose this data because it is useful for real-world applications such as:

- Predicting taxi duration for a trip
- Allocating taxi to zones/regions based on demand.
- Identify whether a toll fee will be paid

Below is the summary of the data attributes as described here

Attribute / Column	Description
VendorID	A code indicating the TPEP provider that provided the record. **1= Creative Mobile Technologies, LLC; 2= VeriFone Inc.**
tpep_pickup_datetime	The date and time when the meter was engaged.
tpep_dropoff_datetime	The date and time when the meter was disengaged.
Passenger_count	The number of passengers in the vehicle. This is a driver-entered value.
Trip_distance	The elapsed trip distance in miles reported by the taximeter.
PULocationID	TLC Taxi Zone in which the taximeter was engaged.
DOLocationID	TLC Taxi Zone in which the taximeter was disengaged.
RateCodeID	The final rate code in effect at the end of the trip. **1= Standard rate** **2=JFK** **3=Newark** **4=Nassau or Westchester** **5=Negotiated fare**

	6=Group ride
Store_and_fwd_flag	This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka "store and forward," because the vehicle did not have a connection to the server. **Y= store and forward trip** **N= not a store and forward trip**
Payment_type	A numeric code signifying how the passenger paid for the trip. **1= Credit card** **2= Cash** **3= No charge** **4= Dispute** **5= Unknown** **6= Voided trip**
Fare_amount	The time-and-distance fare calculated by the meter.
Extra	Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges.
MTA_tax	$0.50 MTA tax that is automatically triggered based on the metered rate in use.
Improvement_surcharge	$0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015.
Tip_amount	Tip amount –This field is automatically populated for credit card tips. Cash tips are not included.
Tolls_amount	Total amount of all tolls paid in trip.
Total_amount	The total amount charged to passengers. Does not include cash tips.

Data Exploration

Joining the 12 files provided resulted in 2, 922, 266 instances in the dataset with no missing values. We have 12 numerical features, 4 categorical features, 2 datetime features and 1 integer. For easy visualization, I converted *VendorID, RateCodeID and payment_type* to their corresponding values.

Summary of some the numerical attributes shows that this data is not perfect and therefore needs some cleaning for example there is no way we can have a negative tip or zero passengers.

	Passenger_count	Trip_distance	Fare_amount	Tip_amount
min	0.0	0.0	-150.00	-2.7
max	9.0	14680110.0	410266.86	650

Data Cleaning and Feature engineering

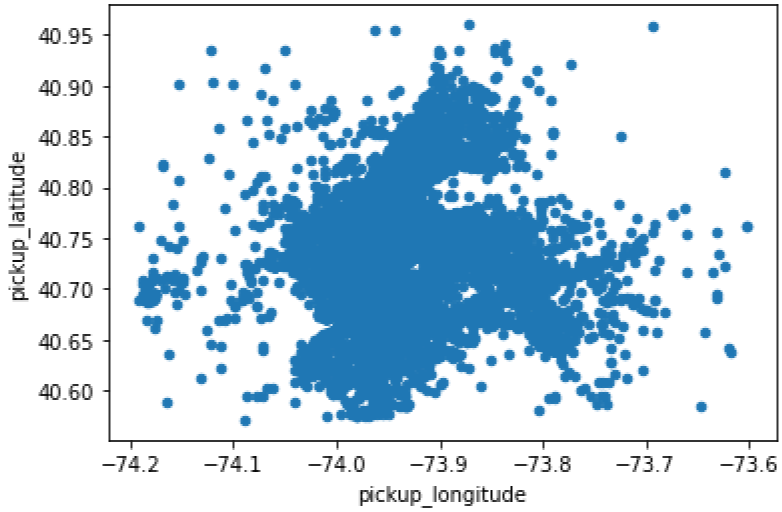
For data preprocessing/cleaning

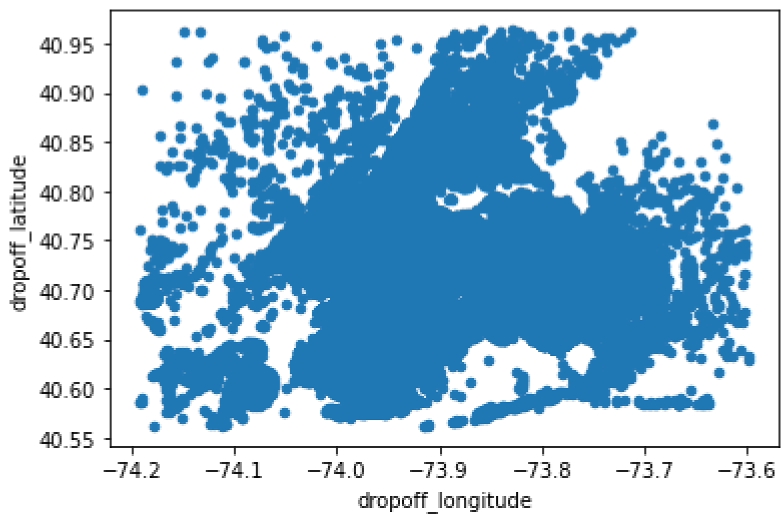
- Charges and trip information that are negative, as well as charges inconsistent with expected values are considered incorrect. In addition, trips with pickup or drop off locations outside a geographic region of interest are removed.
- Only keep trips with valid passenger and distance information.
- Remove trips missing valid pickup or drop off locations.
- Remove trips with invalid Rate code.

For Feature engineering, the following features were added.

- ***duration*** - Length of the trip, in minutes, calculated from the pickup and drop off times.
- ***avespeed*** - Average speed, in mph, calculated from the distance and duration values.
- ***time_of_day*** - This feature represents pick up time as the elapsed time since midnight in decimal hours (e.g. 7:10 am becomes 7.1667). The output is a duration vector with units of hours.
- ***day_of_week*** - This feature is a categorical array indicating the day of the week the trip began, in long format (e.g. 'Monday').
- ***toll_paid*** **–** This feature indicates trips that charged a toll or not.

After removing invalid trip information, we can now visualize some of the features.



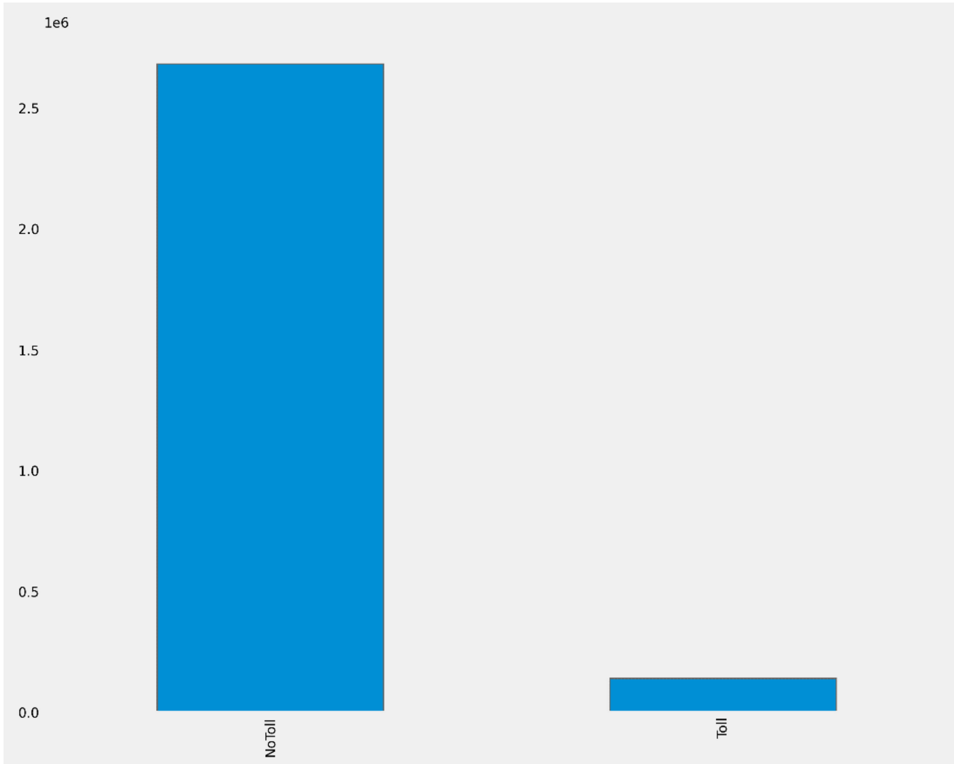


Data Preparation for Machine Learning

Target Variable

We chose the target variable to be ***toll_paid*** as we want to predict the presence or absence of toll(s) for a trip.

	Count	Percentage (%)
No Toll	2682515	0.95
Toll	141186	0.05
	2823701	**100**



The distribution above shows that we have unbalanced classes as 95% of the data belongs to *No Toll* class.

Feature Selection

To avoid using all the 22 columns as features, we tried to find the correlation between each numerical feature and the *tolls_amount*, and the result is as below:

Features	Correlation with tolls_amount
total_amount	0.675407
trip_distance	**0.634744**
fare_amount	0.606455
duration	0.457660
pickup_longitude	**0.434626**
tip_amount	0.431183
ave_speed	**0.412052**
dropoff_longitude	**0.303816**
passenger_count	0.012695
time_of_day	**0.001991**
dropoff_latitude	**-0.054995**
extra	-0.073699
pickup_latitude	**-0.111228**

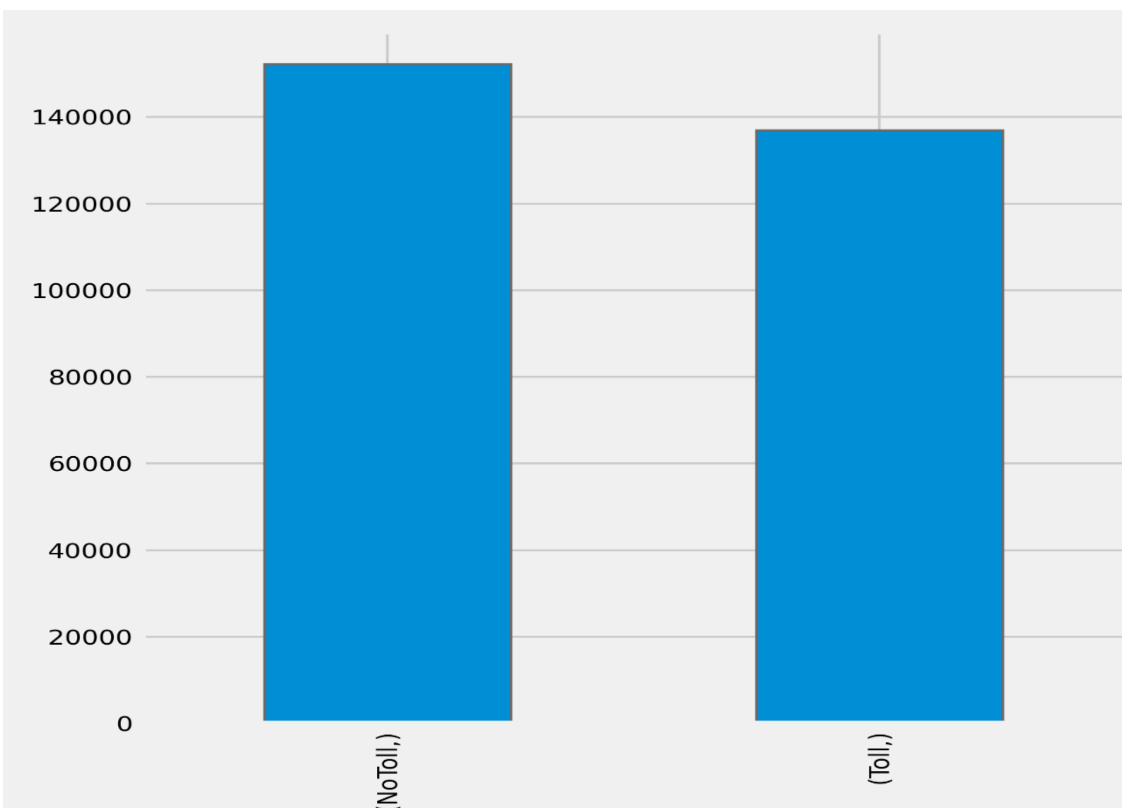
We opted to use the highlighted numerical features and ***day_of_week*** as the only categorical feature, a total of 11 features.

Data Splitting

We first split our data into two (train: 2,738,989 rows and test: 84,712 rows) using sklearn's StratifiedShuffleSplit class to ensure that the distribution of each class is represented accurately.

Keeping the test dataset aside, we employed downsampling technique to address the unbalanced classes with the code snippet and the new distribution is as shown below:

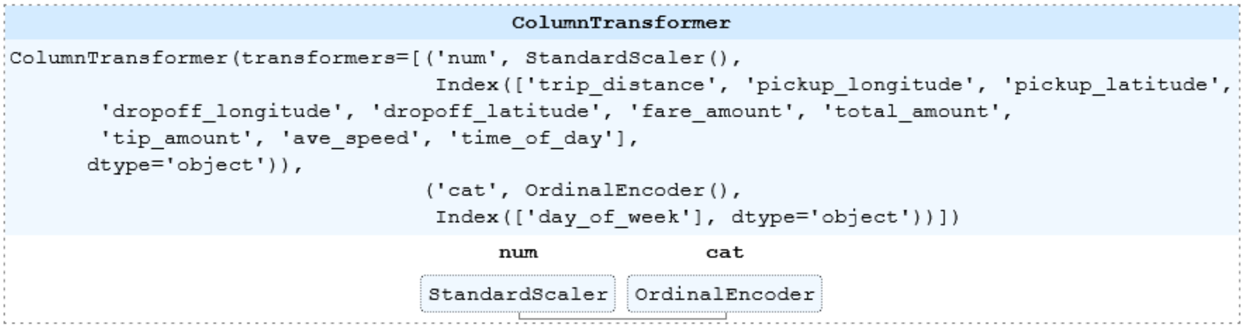
```
from imblearn.under_sampling import RandomUnderSampler

undersample = RandomUnderSampler(sampling_strategy=0.9,
                                 random_state=42,
                                 replacement=False)
sampled = undersample.fit_resample(X_train_val, y_train_val)
X_train_val, y_train_val = sampled
```



We further split the training data into two (train: 231,292 rows and validation: 57,824 rows)

Feature transformation

We applied two transformations on the features: feature scaling using *StandardScaler* for the numerical features and ordinal encoding for the categorical feature. We built below transformation pipeline using scikit-learn's ***ColumnTransformer*** class.

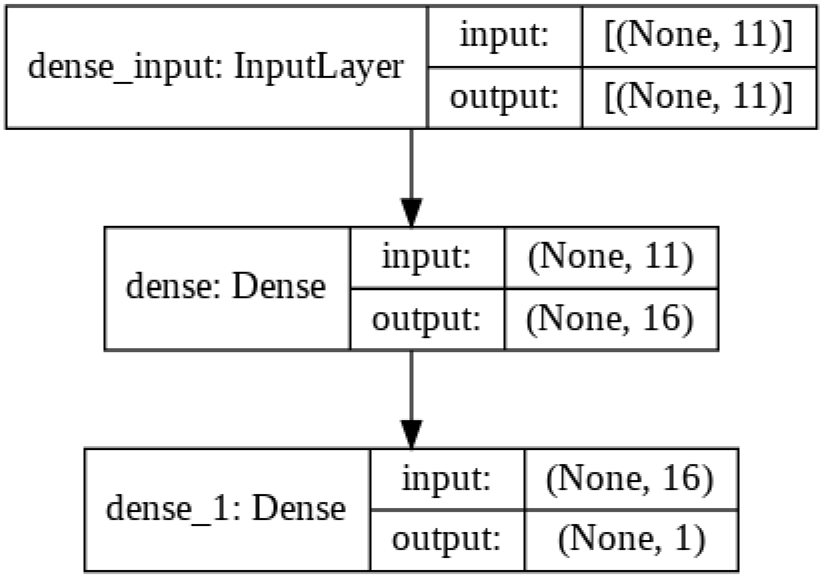


Model Training and Evaluation

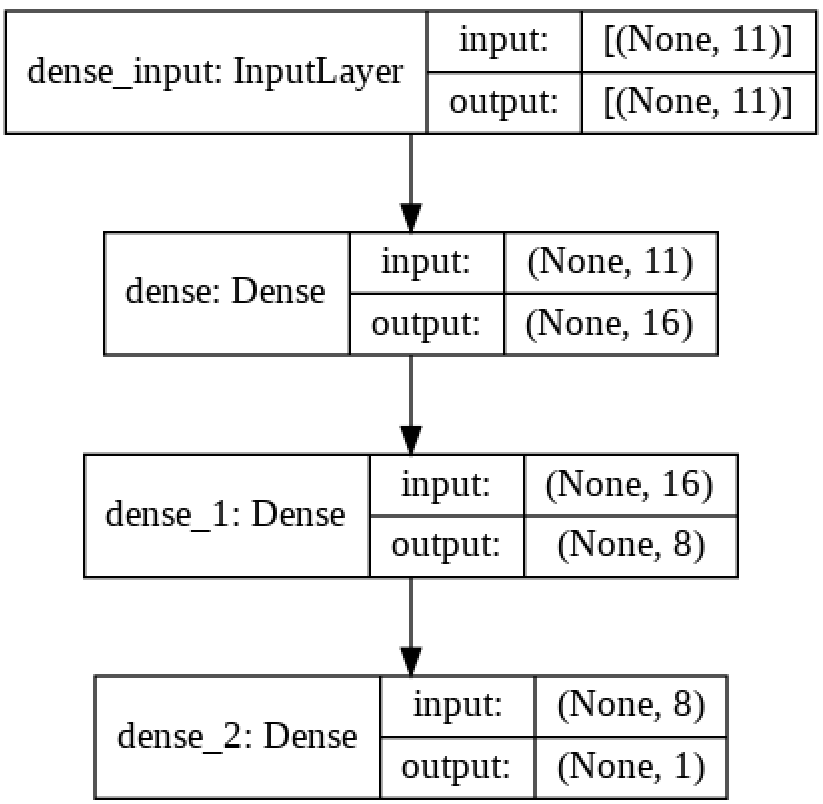
We trained four (4) different classification models which includes 1 classical ML model and 3 deep learning models with different architectures:

- Base Model (Random Forest Classifier)

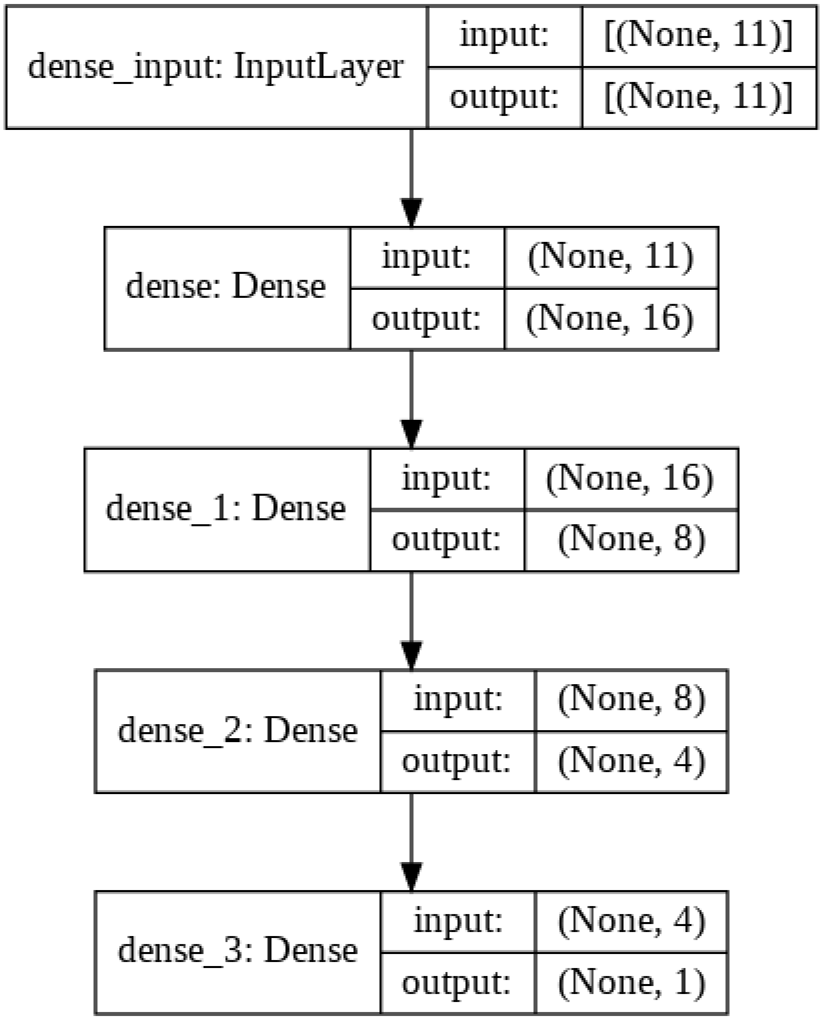
- Model 1



- Model 2



- Model 3

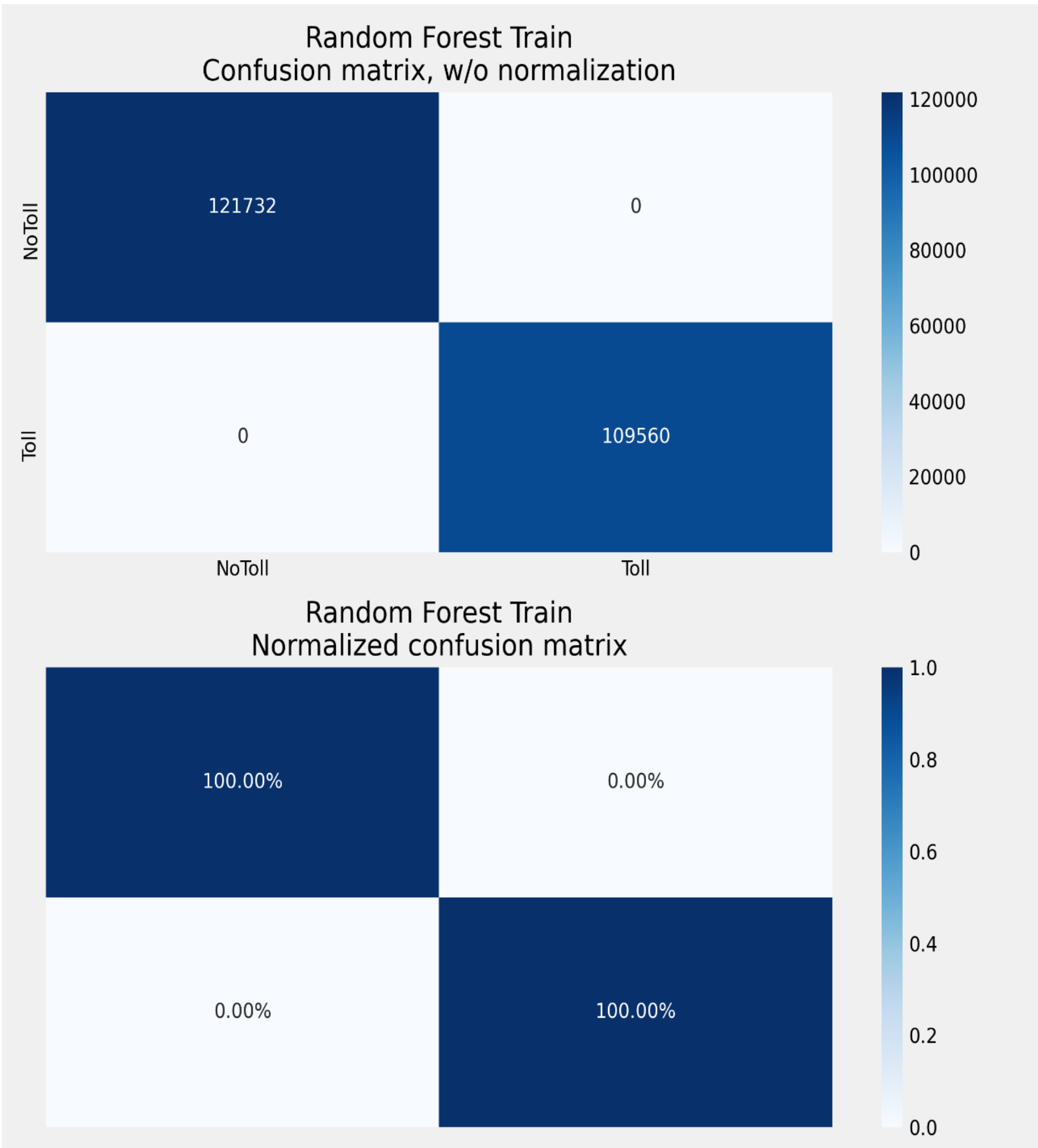


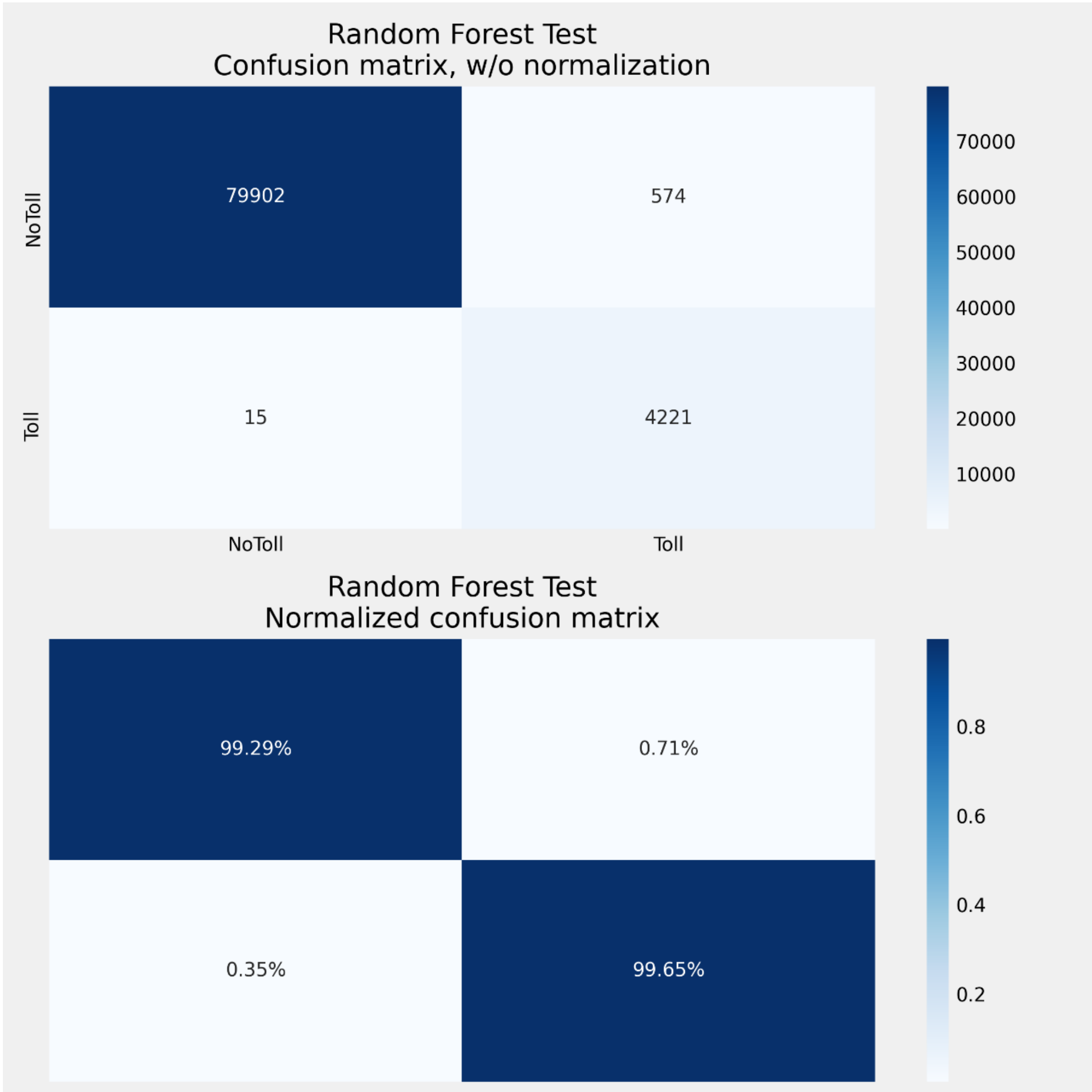
In addition to accuracy, we evaluated the models using below metrics given the fact that we are dealing with a classification problem.

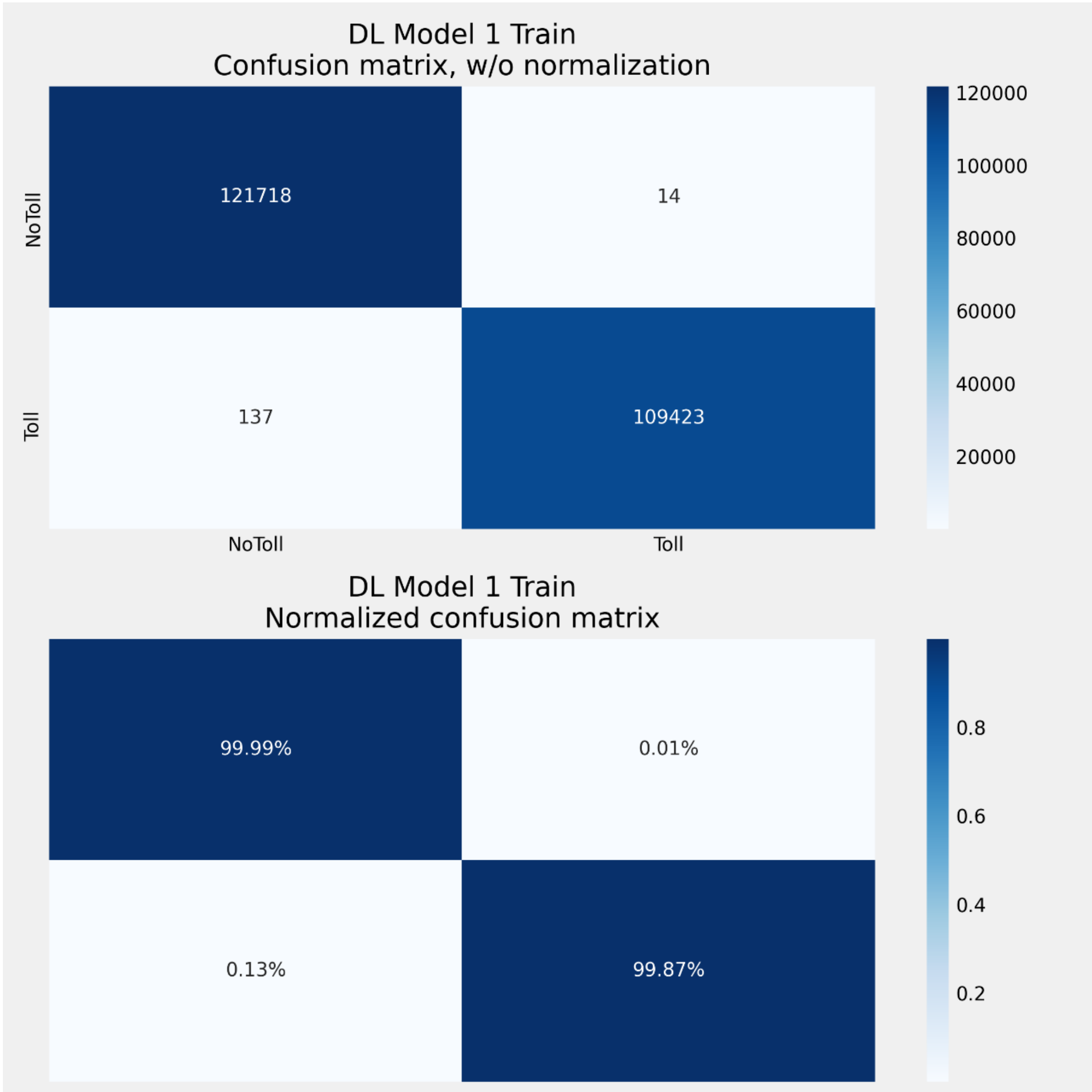
	Random Forest		**Model 1**		**Model 2**		**Model 3**	
	Train	Test	Train	Test	Train	Test	Train	Test
Accuracy	0.999997	0.993047	**0.999191**	**0.999646**	0.999390	0.999669	0.999235	0.999681
Precision	1.000000	0.880292	**0.999744**	**0.995757**	0.999735	0.995759	0.999826	0.996461
Recall	0.999993	0.996459	**0.998549**	**0.997167**	0.998978	0.997639	0.998558	0.997167
F1 Score	0.999996	0.934780	**0.999146**	**0.996461**	0.999356	0.996698	0.999192	0.996814
Auc	0.999996	0.994663	**0.999159**	**0.998472**	0.999370	0.998708	0.999201	0.998490

Key Findings and Insights

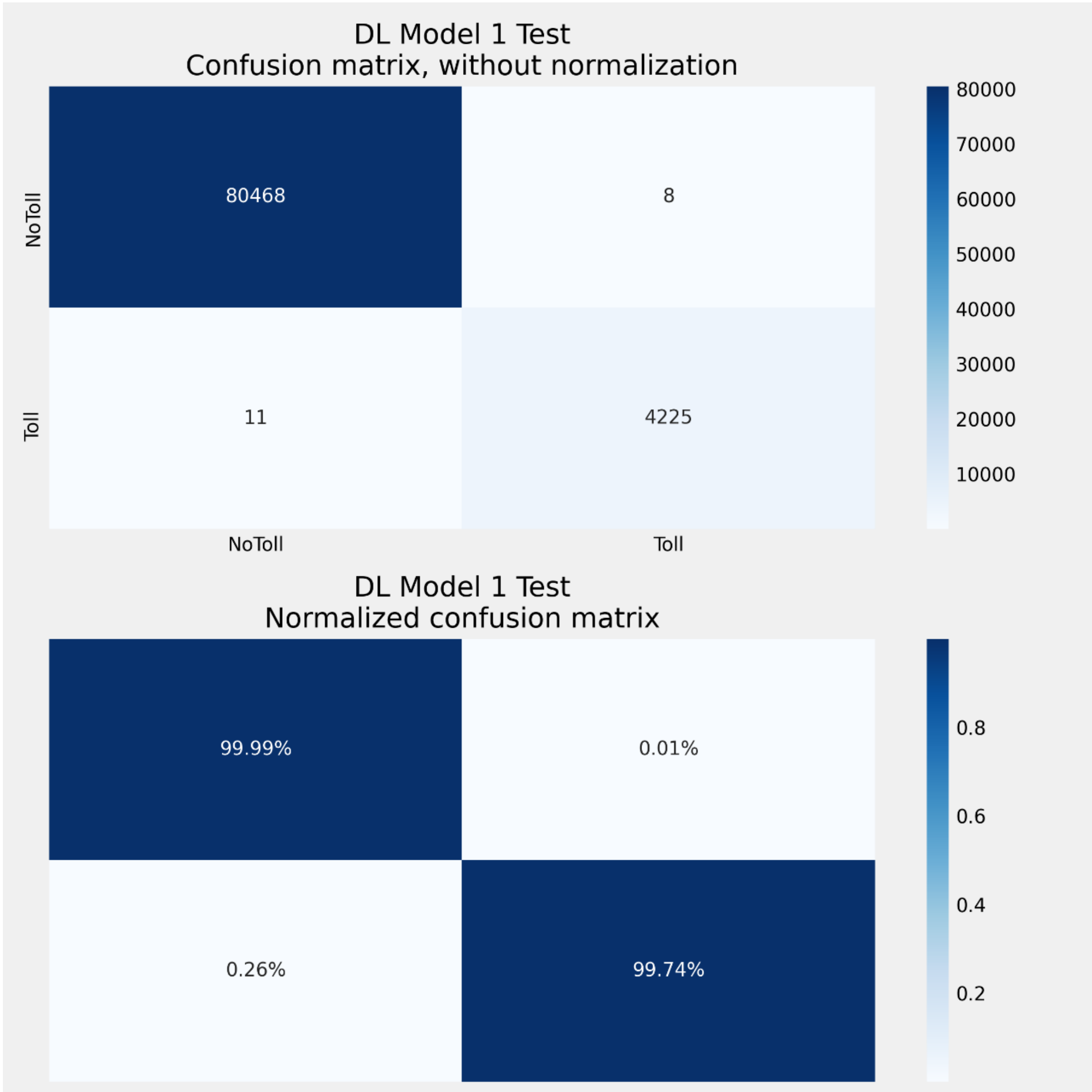
We explored the confusion matrix of all the models to understand the number and percentage of the two classes correctly predicted given the true values for both the training and test data.

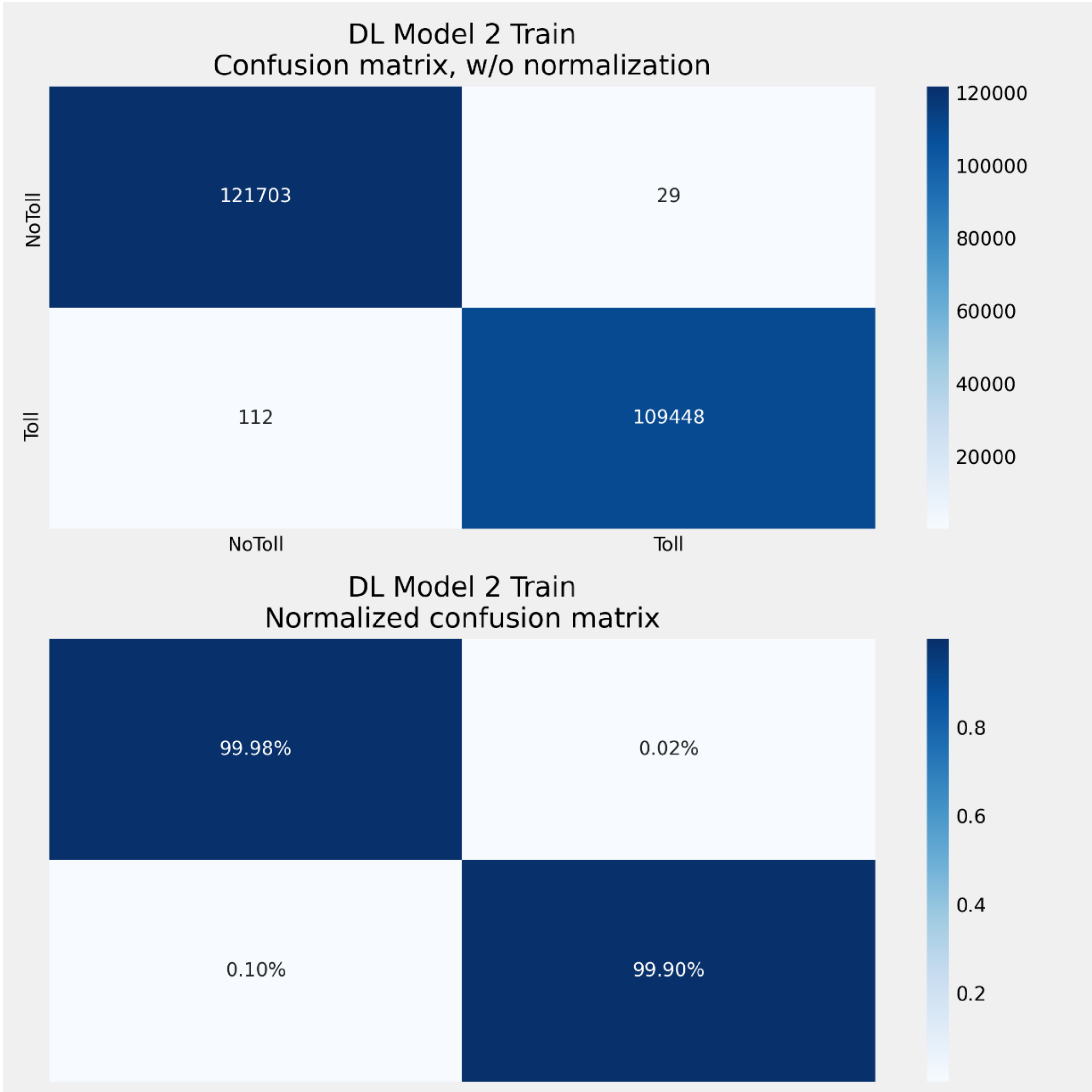


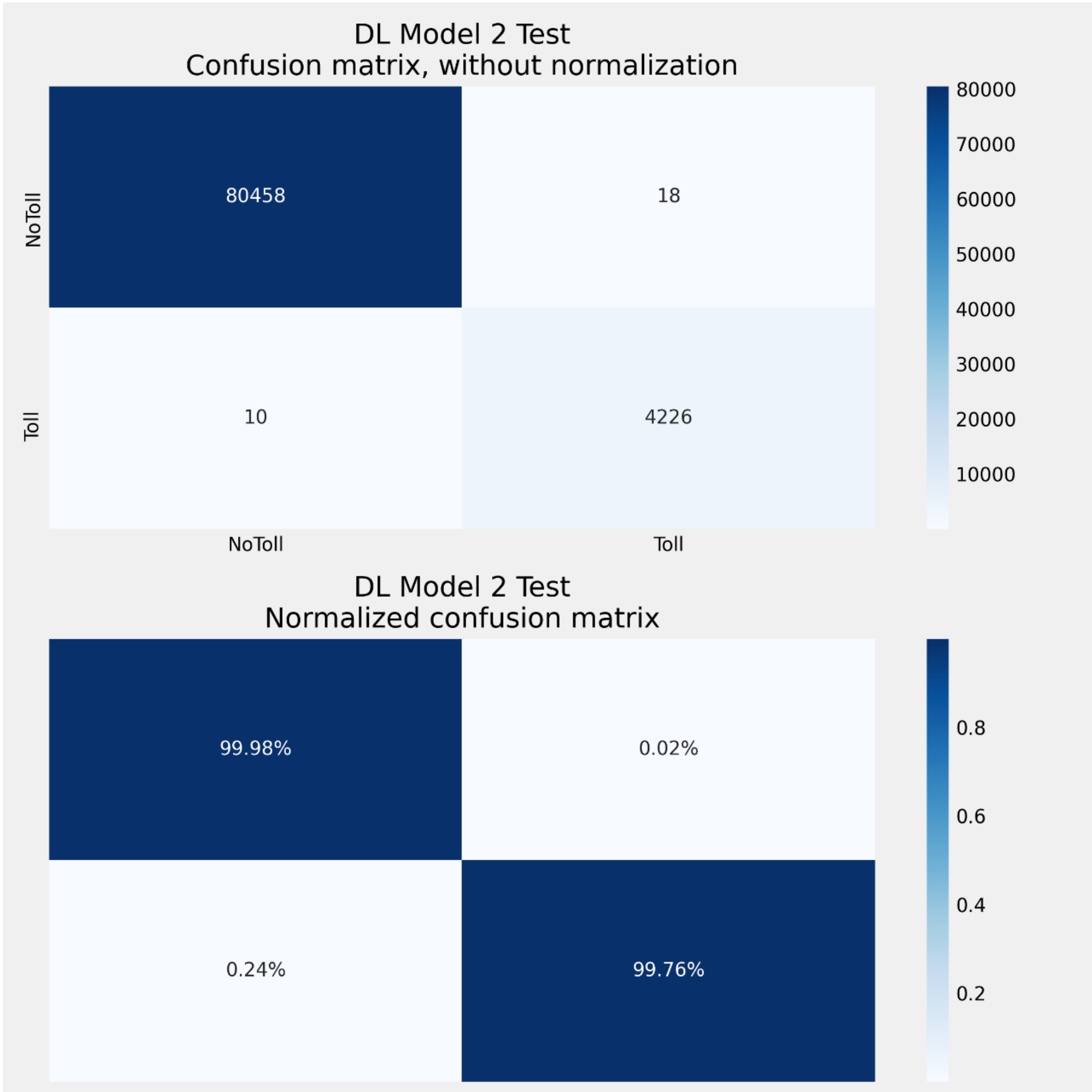

Random Forest Test
Confusion matrix, w/o normalization
NoToll
Toll
79902
574
15
4221
NoToll
Toll
70000
60000
50000
40000
30000
20000
10000
Random Forest Test
Normalized confusion matrix
99.29%
0.71%
0.35%
99.65%
0.8
0.6
0.4
0.2

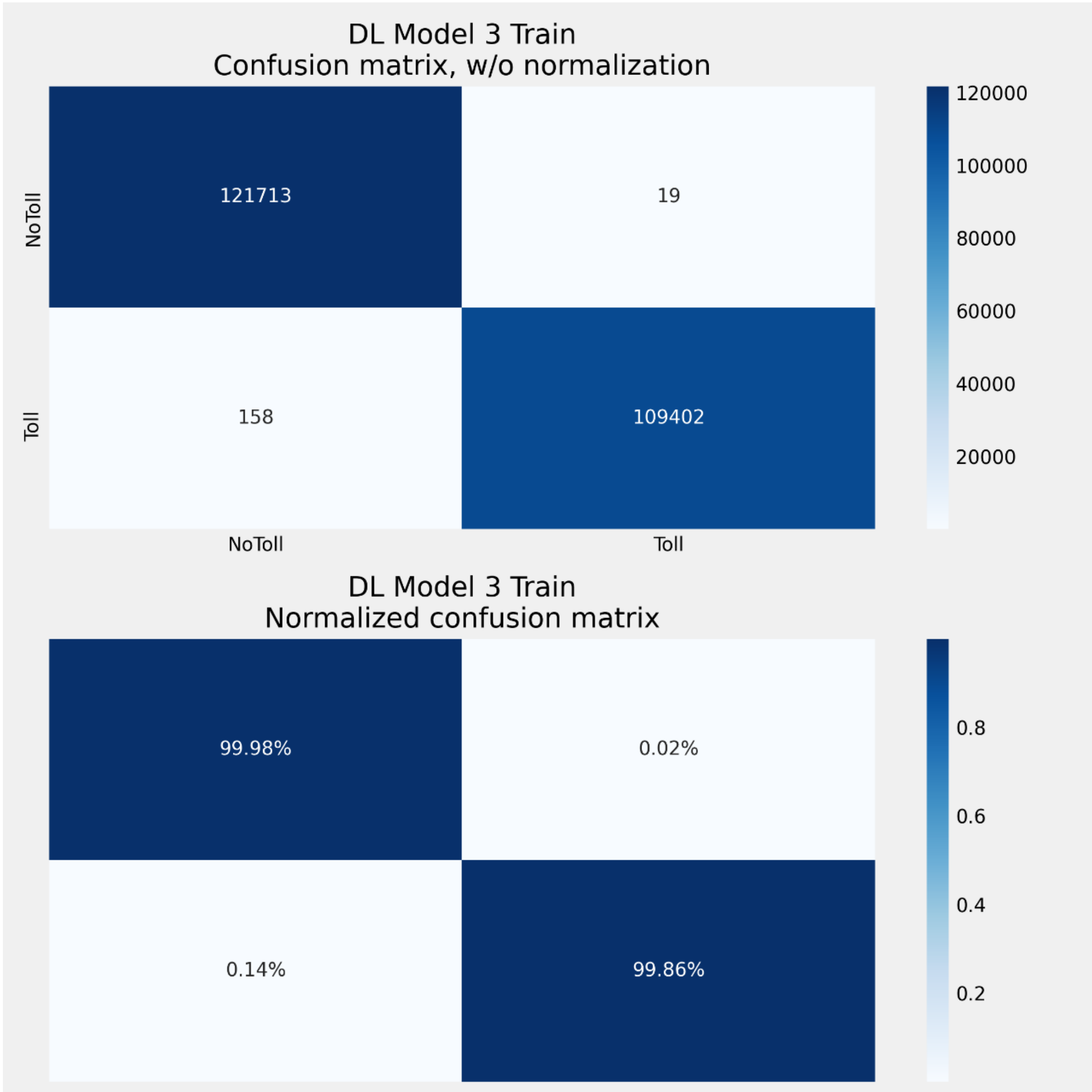


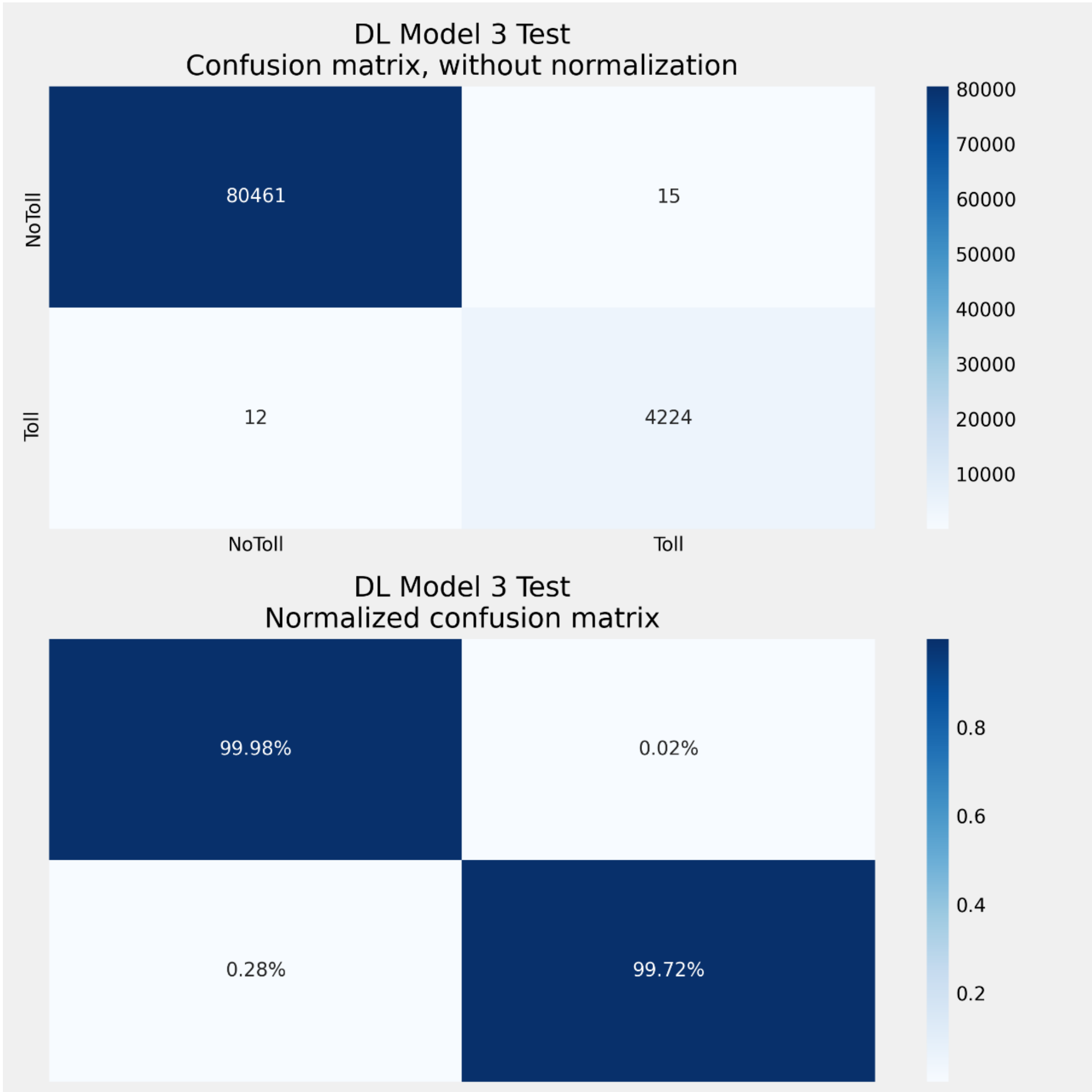
DL Model 1 Train
Confusion matrix, w/o normalization
NoToll
Toll
121718
14
137
109423
NoToll
Toll
120000
100000
80000
60000
40000
20000
DL Model 1 Train
Normalized confusion matrix
99.99%
0.01%
0.13%
99.87%
0.8
0.6
0.4
0.2


DL Model 1 Test
Confusion matrix, without normalization
NoToll
Toll
80468
8
11
4225
NoToll
Toll
80000
70000
60000
50000
40000
30000
20000
10000
DL Model 1 Test
Normalized confusion matrix
99.99%
0.01%
0.26%
99.74%
0.8
0.6
0.4
0.2


DL Model 2 Train
Confusion matrix, w/o normalization
NoToll
Toll
121703
29
112
109448
NoToll
Toll
120000
100000
80000
60000
40000
20000
DL Model 2 Train
Normalized confusion matrix
99.98%
0.02%
0.10%
99.90%
0.8
0.6
0.4
0.2


DL Model 2 Test
Confusion matrix, without normalization
NoToll
Toll
80458
18
10
4226
NoToll
Toll
80000
70000
60000
50000
40000
30000
20000
10000
DL Model 2 Test
Normalized confusion matrix
99.98%
0.02%
0.24%
99.76%
0.8
0.6
0.4
0.2


DL Model 3 Train
Confusion matrix, w/o normalization
NoToll
Toll
121713
19
158
109402
NoToll
Toll
120000
100000
80000
60000
40000
20000
DL Model 3 Train
Normalized confusion matrix
99.98%
0.02%
0.14%
99.86%
0.8
0.6
0.4
0.2



Model Selection

Based on the metrics evaluated and confusion matrix above, we noticed that the Random Forest Classifier overfits the data and the other 3 models seems to fit both the training and test data well.

The model 1 seems to perform better in all metrics and therefore chosen as the best model for our present case.

Next Steps in model improvement

We will explore the misclassifications in detail to understand why they were misclassified and find ways to address them.

Summary

This model will be very useful in predicting Toll or No Toll for every trip and this will help the business to estimate as correct as possible the total amount a customer will be charged. In addition, it will improve transparency as customers will be able to see exactly what they are being charged.